CLAIMS

1. A method for verifying relevance between terms and Web site contents, the method comprising:

retrieving site contents from a bid URL;

formulating expanded term(s) semantically and/or contextually related to bid term(s),

generating content similarity and expanded similarity measurements from respective combinations of the bid term(s), the site contents, and the expanded terms, the similarity measurements indicating relatedness between respective ones of the bid term(s), site contents, and/or expanded terms;

calculating category similarity measurements between the expanded terms and the site contents in view of a similarity classifier, the similarity classifier having been trained from mined web site content associated with directory data;

calculating a confidence value from combined ones of multiple similarity measurements, the combined ones comprising content, expanded, and category similarity measurements, the confidence value providing an objective measure of relevance between the bid term(s) and the site contents.

2. A method as recited in claim 1, wherein the similarity classifier is based on a statistical n-gram based naïve Bayesian (N-Gram), a naïve Bayesian (NB), support vector machine (SVM), a nearest neighbor (KNN), a decision tree, a co-training, or a boosting classification model.

3. A method as recited in claim 1, wherein formulating the expanded terms further comprises generating term clusters from term vectors based on calculated term similarity, the term vectors being generated from historical queries, each historical query having a high frequency of occurrence, the term clusters comprising the expanded terms.

4. A method as recited in claim 1, wherein generating the content similarity measurements further comprise generating respective term vectors from the bid term(s) and the site contents, and calculating similarity between the respective term vectors to determine direct similarity between the bid term(s) and the site contents.

5. A method as recited in claim 1, wherein generating the expanded similarity measurements further comprises:

generating respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating similarity between the respective term vectors to determine the expanded similarity measurements between the bid term(s) and the site contents.

6. A method as recited in claim 1, wherein generating the category similarity measurements further comprises:

extracting features from Web site content associated with the directory data, the features comprising a combination of title, metadata, body, hypertext link(s), visual feature(s), and/or summarization by page layout analysis information;

reducing dimensionality of the features via feature selection;

categorizing the features via a classifier model to generate the similarity classifier;

generating respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating similarity between the respective term vectors as a function of the similarity classifier to determine the category similarity measurements.

7. A method as recited in claim 1, wherein calculating the confidence value further comprises:

training a combined relevance classifier with data of the form <term(s), Web site content, accept/reject> in view of an accept/reject threshold;

generating relevance verification similarity measurement (RSVM) feature vectors from the content, expanded, and category similarity measurements; and

mapping multiple scores from the RSVM feature vectors to the confidence value via the combined relevance classifier.

8. A method as recited in claim 1, wherein the method further comprises:

caching the bid term(s) and bid URL into a bidding database;

responsive to receipt of an search query, determining if terms of the search query are relevant to the bid term(s) in view of a possibility that the terms of the search query may not exactly match the bid term(s); and

if the term(s) of search query are determined to be relevant to the bid term(s), communicating the bid URL to the end-user.

9. A method as recited in claim 1, wherein the method further comprises:

determining proper name similarity measurements from the bid term(s) and site contents, the proper name similarity measurements indicating relatedness between any proper name(s) detected in the bid term(s) and the site contents in view a set of proper names; and

wherein the combined ones of multiple similarity measurements comprise the proper name similarity measurements.

10. A method as recited in claim 9, wherein determining the proper name similarity measurements further comprises:

responsive to detecting a proper name in the bid term(s) and/or the site contents, calculating a proper name similarity score as:

Prop_Sim(*term, site contents*),

whrerein Prop_Sim(*term, site contents*) equals: one (1) when a *term* contains a proper name *P,* and *site contents* contains a conformable proper name *Q*; zero (0)

when a *term* contains a proper name *P*, and *site contents* contains only unconformable proper name(s); or, zero-point-five (0.5).

11. A method as recited in claim 1, wherein the method further comprises:

determining that the confidence value is relatively low; and

responsive to the determining, identifying one or more other terms that are semantically and/or contextually related to the bid URL.

12. A method as recited in claim 11, wherein identifying further comprises:

generating a set of term clusters from term vectors based on calculated term similarity, the term vectors being generated from search engine results of submitted historical queries, each historical query having a relatively low frequency of occurrence as compared to other query terms in a query log; and

evaluating the site contents in view of term(s) specified by the term clusters to identify one or more semantically and/or contextually related terms, the terms being the one or more other terms.

13. A computer-readable medium comprising computer-executable instructions for verifying relevance between terms and Web site contents, the computer-executable instructions comprising instructions for:

retrieving site contents from a bid URL;

formulating expanded term(s) semantically and/or contextually related to bid term(s),

generating content similarity and expanded similarity measurements from respective combinations of the bid term(s), the site contents, and the expanded terms, the similarity measurements indicating relatedness between respective ones of the bid term(s), site contents, and/or expanded terms;

calculating category similarity measurements between the expanded terms and the site contents in view of a similarity classifier, the similarity classifier having been trained from mined web site content associated with directory data;

calculating a confidence value from combined ones of multiple similarity measurements, the combined ones comprising content, expanded, and category similarity measurements, the confidence value providing an objective measure of relevance between the bid term(s) and the site contents.

14. A computer-readable medium as recited in claim 13, wherein the similarity classifier is based on a statistical n-gram based naïve Bayesian (N-Gram), a naïve Bayesian (NB), support vector machine (SVM), a nearest neighbor (KNN), a decision tree, a co-training, or a boosting classification model.

15. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions for formulating the expanded terms further comprise instructions for generating term clusters from term vectors based on calculated term similarity, the term vectors being generated from historical queries, each historical query having a high frequency of occurrence, the term clusters comprising the expanded terms.

16. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions for generating the content similarity measurements further comprise instructions for generating respective term vectors from the bid term(s) and the site contents, and calculating similarity between the respective term vectors to determine direct similarity between the bid term(s) and the site contents.

17. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions for generating the expanded similarity measurements further comprise instructions for:

generating respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating similarity between the respective term vectors to determine the expanded similarity measurements between the bid term(s) and the site contents.

18. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions for generating the category similarity measurements further comprise instructions for:

extracting features from Web site content associated with the directory data, the features comprising a combination of title, metadata, body, hypertext link(s), visual feature(s), and/or summarization by page layout analysis information;

reducing dimensionality of the features via feature selection;

categorizing the features via a classifier model to generate the similarity classifier;

generating respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating similarity between the respective term vectors as a function of the similarity classifier to determine the category similarity measurements.

19. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions for calculating the confidence value further comprise instructions for:

training a combined relevance classifier with data of the form <term(s), Web site content, accept/reject> in view of an accept/reject threshold;

generating relevance verification similarity measurement (RSVM) feature vectors from the content, expanded, and category similarity measurements; and

mapping multiple scores from the RSVM feature vectors to the confidence value via the combined relevance classifier.

20. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions further comprise instructions for:

caching the bid term(s) and bid URL into a bidding database;

responsive to receipt of an search query, determining if terms of the search query are relevant to the bid term(s) in view of a possibility that the terms of the search query may not exactly match the bid term(s); and

if the term(s) of search query are determined to be relevant to the bid term(s), communicating the bid URL to the end-user.

21. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions further comprise instructions for:

determining proper name similarity measurements from the bid term(s) and site contents, the proper name similarity measurements indicating relatedness between any proper name(s) detected in the bid term(s) and the site contents in view a set of proper names; and

wherein the combined ones of multiple similarity measurements comprise the proper name similarity measurements.

22. A computer-readable medium as recited in claim 21, wherein the computer-executable instructions for determining the proper name similarity measurements further comprise instructions for:

responsive to detecting a proper name in the bid term(s) and/or the site contents, calculating a proper name similarity score as:

Prop_Sim(*term, site contents*) and

wherein Prop_Sim(*term, site contents*) equals: one (1) when a *term* contains a proper name *P,* and *site contents* contains a conformable proper name *Q*; zero (0) when a *term* contains a proper name *P*, and *site contents* contains only unconformable proper name(s); or, zero-point-five (0.5).

23. A computer-readable medium as recited in claim 13, wherein the computer-executable instructions further comprise instructions for:

determining that the confidence value is relatively low; and

responsive to the determining, identifying one or more other terms that are semantically and/or contextually related to the bid URL.

24. A computer-readable medium as recited in claim 23, wherein the computer-executable instructions for identifying further comprise instructions for:

generating a set of term clusters from term vectors based on calculated term similarity, the term vectors being generated from search engine results of submitted historical queries, each historical query having a relatively low frequency of occurrence as compared to other query terms in a query log; and

evaluating the site contents in view of term(s) specified by the term clusters to identify one or more semantically and/or contextually related terms, the terms being the one or more other terms.

25. A computing device for verifying relevance between terms and Web site contents, the computing device comprising:

a processor; and

a memory coupled to the processor, the memory comprising computer-program instructions executable by the processor for:

retrieving site contents from a bid URL;

formulating expanded term(s) semantically and/or contextually related to bid term(s),

generating content similarity and expanded similarity measurements from respective combinations of the bid term(s), the site contents, and the expanded terms, the similarity measurements indicating relatedness between respective ones of the bid term(s), site contents, and/or expanded terms;

calculating category similarity measurements between the expanded terms and the site contents in view of a similarity classifier, the similarity classifier having been trained from mined web site content associated with directory data;

calculating a confidence value from combined ones of multiple similarity measurements, the combined ones comprising content, expanded, and category similarity measurements, the confidence value providing an objective measure of relevance between the bid term(s) and the site contents.

26. A computing device as recited in claim 25, wherein the similarity classifier is based on a statistical n-gram based naïve Bayesian (N-Gram), a naïve Bayesian (NB), support vector machine (SVM), a nearest neighbor (KNN), a decision tree, a co-training, or a boosting classification model.

27. A computing device as recited in claim 25, wherein the computer-executable instructions for formulating the expanded terms further comprise instructions for generating term clusters from term vectors based on calculated term similarity, the term vectors being generated from historical queries, each historical query having a high frequency of occurrence, the term clusters comprising the expanded terms.

28. A computing device as recited in claim 25, wherein the computer-executable instructions for generating the content similarity measurements further comprise instructions for generating respective term vectors from the bid term(s) and the site contents, and calculating similarity between the respective term vectors to determine direct similarity between the bid term(s) and the site contents.

29. A computing device as recited in claim 25, wherein the computer-executable instructions for generating the expanded similarity measurements further comprise instructions for:

generating respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating similarity between the respective term vectors to determine the expanded similarity measurements between the bid term(s) and the site contents.

30. A computing device as recited in claim 25, wherein the computer-executable instructions for generating the category similarity measurements further comprise instructions for:

extracting features from Web site content associated with the directory data, the features comprising a combination of title, metadata, body, hypertext link(s), visual feature(s), and/or summarization by page layout analysis information;

reducing dimensionality of the features via feature selection;

categorizing the features via a classifier model to generate the similarity classifier;

generating respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating similarity between the respective term vectors as a function of the similarity classifier to determine the category similarity measurements.

31. A computing device as recited in claim 25, wherein the computer-executable instructions for calculating the confidence value further comprise instructions for:

training a combined relevance classifier with data of the form <term(s), Web site content, accept/reject> in view of an accept/reject threshold;

generating relevance verification similarity measurement (RSVM) feature vectors from the content, expanded, and category similarity measurements; and

mapping multiple scores from the RSVM feature vectors to the confidence value via the combined relevance classifier.

32. A computing device as recited in claim 25, wherein the computer-executable instructions further comprise instructions for:

determining proper name similarity measurements from the bid term(s) and site contents, the proper name similarity measurements indicating relatedness between any proper name(s) detected in the bid term(s) and the site contents in view a set of proper names; and

wherein the combined ones of multiple similarity measurements comprise the proper name similarity measurements.

33. A computing device as recited in claim 32, wherein the computer-executable instructions for determining the proper name similarity measurements further comprise instructions for:

responsive to detecting a proper name in the bid term(s) and/or the site contents, calculating a proper name similarity score as:

Prop_Sim(*term, site contents*) and

wherein Prop_Sim(*term, site contents*) equals: one (1) when a *term* contains a proper name *P,* and *site contents* contains a conformable proper name *Q*; zero (0) when a *term* contains a proper name *P*, and *site contents* contains only unconformable proper name(s); or, zero-point-five (0.5).

34. A computing device as recited in claim 25, wherein the computer-executable instructions further comprise instructions for:

determining that the confidence value is relatively low; and

responsive to the determining, identifying one or more other terms that are semantically and/or contextually related to the bid URL.

35. A computing device as recited in claim 34, wherein the computer-executable instructions for identifying further comprise instructions for:

generating a set of term clusters from term vectors based on calculated term similarity, the term vectors being generated from search engine results of submitted historical queries, each historical query having a relatively low frequency of occurrence as compared to other query terms in a query log; and

evaluating the site contents in view of term(s) specified by the term clusters to identify one or more semantically and/or contextually related terms, the terms being the one or more other terms.

36. A computing device for verifying relevance between terms and Web site contents, the computing device comprising:

retrieving means to obtain site contents from a bid URL;

formulating means to identify expanded term(s) semantically and/or contextually related to bid term(s),

generating means to create content similarity and expanded similarity measurements from respective combinations of the bid term(s), the site contents, and the expanded terms, the similarity measurements indicating relatedness between respective ones of the bid term(s), site contents, and/or expanded terms;

calculating means to determine category similarity measurements between the expanded terms and the site contents in view of a similarity classifier, the similarity classifier having been trained from mined web site content associated with directory data;

calculating means to generate a confidence value from combined ones of multiple similarity measurements, the combined ones comprising content, expanded, and category similarity measurements, the confidence value providing an objective measure of relevance between the bid term(s) and the site contents.

37. A computing device as recited in claim 36, wherein the computer formulating means further comprise generating means to create term clusters from term vectors based on calculated term similarity, the term vectors being generated from historical queries, each historical query having a high frequency of occurrence, the term clusters comprising the expanded terms.

38. A computing device as recited in claim 36, wherein the generating means further comprise creating means to generate respective term vectors from the bid term(s) and the site contents, and calculating similarity between the respective term vectors to determine direct similarity between the bid term(s) and the site contents.

39. A computing device as recited in claim 36, wherein the generating means further comprise:

creating means to generate respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating means to determine similarity between the respective term vectors to determine the expanded similarity measurements between the bid term(s) and the site contents.

40. A computing device as recited in claim 36, wherein the generating means further comprise:

extracting means to obtain features from Web site content associated with the directory data, the features comprising a combination of title, metadata, body, hypertext link(s), visual feature(s), and/or summarization by page layout analysis information;

reducing means to lessen dimensionality of the features via feature selection;

categorizing means to organize the features via a classifier model to generate the similarity classifier;

generating means to create respective term vectors from the bid term(s), the site contents, and the expanded terms; and

calculating means to identify similarity between the respective term vectors as a function of the similarity classifier to determine the category similarity measurements.

41. A computing device as recited in claim 36, wherein the calculating means further comprise:

training means to train a combined relevance classifier with data of the form <term(s), Web site content, accept/reject> in view of an accept/reject threshold;

generating means to generate relevance verification similarity measurement (RSVM) feature vectors from the content, expanded, and category similarity measurements; and

mapping means to correlate multiple scores from the RSVM feature vectors to the confidence value via the combined relevance classifier.

42. A computing device as recited in claim 36, wherein the computing device further comprises:

determining means to determine proper name similarity measurements from the bid term(s) and site contents, the proper name similarity measurements indicating relatedness between any proper name(s) detected in the bid term(s) and the site contents in view a set of proper names; and

wherein the combined ones of multiple similarity measurements comprise the proper name similarity measurements.

43. A computing device as recited in claim 42, wherein the determining means to determine the proper name similarity measurements further comprise responsive to detecting a proper name in the bid term(s) and/or the site contents, calculating means to calculate a proper name similarity score.

44. A computing device as recited in claim 36, wherein the computing device further comprises:

determining means to determine that the confidence value is relatively low; and

responsive to the determining, identifying means to identify one or more other terms that are semantically and/or contextually related to the bid URL.

45. A computing device as recited in 44, wherein the identifying means further comprise:

generating means to generate a set of term clusters from term vectors based on calculated term similarity, the term vectors being generated from search engine results of submitted historical queries, each historical query having a relatively low frequency of occurrence as compared to other query terms in a query log; and

evaluating means to evaluate the site contents in view of term(s) specified by the term clusters to identify one or more semantically and/or contextually related terms, the terms being the one or more other terms.